"Synthèse de commentaires sportifs:intégration d'une annotation prosodique à deux niveauxdans un synthétiseur HMM"

Brognaux, Sandrine ; Picart, Benjamin ; Drugman, Thomas

Abstract

This paper proposes a new prosody annotation protocol specific to live sports commentaries. Two levels of annotation are defined with HMM-based speech synthesis in view. Local labels are assigned to all syllables and refer to accentual phenomena. Global labels classify sequences of words into five distinct sub-genres, defined in terms of valence and arousal. Our analysis shows that the labels are both related to a specific function and characterized by a distinct acoustic realization. The consideration of these constraints should allow for an automatic prediction of the labels both from the text or from the speech signal. The integration of this new annotation protocol within HMM-based speech synthesis shows promising results.

Document type : *Article de périodique (Journal article)*

Variations phonétiques: Impact de la situation de communication

Sandrine Brognaux[1,2], Thomas Drugman[2]
[1]Cental, ICTEAM (Université catholique de Louvain), Belgium
[2]TCTS Lab (Université de Mons), Belgium
<sandrine.brognaux@uclouvain.be, thomas.drugman@umons.ac.bee>

Résumé

While speech synthesis research is now focussing on the generation of various speaking styles or emotions, few studies have considered the possibility of including phonetic variations according to the communicative situation of the targeted speech (sports commentaries, TV news, etc.). This paper offers a phonetic analysis of large French corpora to assess the influence exerted by three situational 'traits': read/spontaneous, media/non-media and expressive/non-expressive. It shows that some variations, like elision, tend to be more frequent in spontaneous and non-media speech, conversely to liaisons which appear more often in read and media speech. Interestingly, no phonetic variation draws a clearcut distinction between expressive and non-expressive speech. Finally, a prosodic analysis indicates that the phonetic variations are not directly correlated with the rhythmic features of their corresponding situational 'trait'.

Mots clés : *phonétique, rythme, style de parole, synthèse de la parole*

1. Introduction

La synthèse vocale ayant atteint, ces dernières décennies, un bon niveau de qualité et d'intelligibilité pour la génération de parole neutre lue, la recherche se tourne à présent vers la synthèse de différents styles de parole et émotions. La plupart des études sur ce sujet se concentrent exclusivement sur la modification de la prosodie ou de la qualité vocale (voir (Yamagishi, Onishi, Masuko, & Kobayashi, 2005 ; Tsuzuki et al., 2004)). Il est étonnant de constater que les modifications phonétiques potentielles de la phrase à synthétiser ont souvent été négligées. Une rare exception est présentée par Roekhaut, Goldman, et Simon (2010), qui modifient la prononciation des schwas finaux selon la situation de communication (nouvelles radiophoniques, conversation, etc.).

Cette question est particulièrement pertinente en français, langue caractérisée par un nombre élevé de variantes phonétiques. L'élision du schwa et la liaison sont les deux modifications les plus fréquentes. De nombreuses études linguistiques se sont penchées sur les modalités de réalisation de ces phénomènes (p. ex. (Fougeron, Goldman, & Frauenfelder, 2001 ; Hansen, 1991 ; Burki, Ernestus, Gendrot, Fougeron, & Frauenfelder, 2011 ; Hambye, 2005)). Elles montrent que la réalisation des différentes variantes peut s'expliquer par différents facteurs : la morpho-syntaxe (Hambye, 2005), le débit de parole (Lacheret-Dujour, 1991 ; Burki et al., 2011 ; Fougeron, Goldman, Dart, Gulat, & Jeager, 2001), la fréquence lexicale (Fougeron, Goldman, & Frauenfelder, 2001 ; Fougeron, Goldman, Dart, et al., 2001), la probabilité du mot (Jurafsky, Bell, Gregory, & Raymond, 2001), le degré d'articulation (Picart, Drugman, & Dutoit, 2010), l'origine du locuteur (Hambye, 2005 ; Martinet, 1971), l'âge du locuteur (Hambye, 2005), etc. Peu d'études linguistiques ont cependant analysé dans quelle mesure ces variations phonétiques sont influencées par la situation de communication (SC), parfois également appelée 'phonogenre' (Goldman, Auchlin, & Simon, 2009 ; Pršir, Goldman, & Auchlin, 2013). Cependant, la potentielle interaction entre ces deux niveaux est très largement reconnue (Burki et al., 2011 ; Simon, Auchlin, Avanzi, & Goldman, 2009) et l'influence de la SC sur la prosodie a été analysée dans de nombreuses études (Roekhaut et al., 2010 ; Simon et al., 2009 ; Pršir et al., 2013). En ce qui concerne les variations phonétiques, seule l'influence du 'trait' lu / spontané a attiré l'intérêt des chercheurs (Hansen, 1991 ; Fougeron, Goldman, & Frauenfelder, 2001 ; Lucci, 1983).

La majorité des synthétiseurs vocaux intègrent des variations phonétiques de base. Cependant, ils sont entraînés à générer une prononciation correspondant à de la parole neutre lue. En ce qui concerne les variations optionnelles, la variante la plus probable est généralement produite, et ce indépendamment de la SC visée. Tandis que la recherche se concentre à présent sur la génération de parole expressive (Yamagishi et al., 2005 ; Tsuzuki et al., 2004) ou média (p.ex les commentaires sportifs (Picart, Brognaux, & Drugman, 2013)), la nécessité d'une étude de l'influence de ces 'traits' situationnels (tels que définis ci-dessous) se fait ressentir.

Notre étude propose une analyse de l'influence exercée par la SC sur la réalisation phonétique. Ces SC étant difficilement échelonnables sur un seul axe, elles sont ici définies selon trois 'traits' binaires (média / non média, expressif / non expressif et lu / spontané) dénommés

'traits' situationnels dans la suite de cet article. L'objectif principal de notre étude est d'offrir une description éclairée des caractéristiques phonétiques de chaque 'trait' afin de mettre en exergue les variations qui devraient être considérées lors de la synthèse de parole avec une certaine visée communicative. Notre analyse présente l'avantage de s'appuyer sur un corpus en français de taille importante comprenant 32 locuteurs et 10 situations de communication (commentaires sportifs, discours politiques, etc.). L'étude de la réalisation phonétique repose sur une stratégie exploitant des techniques de traitement automatique du langage (TAL). Dans un second temps, notre analyse se concentre sur les caractéristiques rythmiques des différents 'traits', le rythme étant considéré comme l'un des corrélats prosodiques des variations phonétiques (Lacheret-Dujour, 1991 ; Fougeron, Goldman, Dart, et al., 2001).

Notre article s'organise comme suit. La Section 2 présente le corpus et son annotation. La méthodologie développée dans le cadre de notre étude est détaillée en Section 3. L'analyse du corpus, tant phonétique que rythmique, est décrite et discutée en Section 4. Enfin, la Section 5 conclut l'article et présente les perspectives de recherche.

2. Description du corpus

Notre corpus est une version étendue de C-PROM (Avanzi, Simon, Goldman, & Auchlin, 2010) enrichie de sous-corpus utilisés en synthèse vocale. Une attention toute particulière est ici accordée aux commentaires sportifs (Brognaux, Picart, & Drugman, 2013) qui ont également été ajoutés au corpus. La phonétisation du son a été effectuée automatiquement puis vérifiée manuellement. L'entièreté du corpus a ensuite été phonétiquement alignée avec EasyAlign (Goldman, 2011) ou Train &Align (Brognaux, Roekhaut, Drugman, & Beaufort, 2012).

Le corpus a une durée totale d'environ 300 minutes, comprend 32 locuteurs (français, belges et suisses) et 10 sous-corpus correspondant à différentes SC (interview, discours politique, etc.). Chaque situation contient 2 à 7 locuteurs et est définie selon trois 'traits' situationnels binaires : média, lu et expressif. Expressif s'entend ici dans le sens d'une implication émotive audible du locuteur (excitation, colère, joie, etc.), qu'elle soit actée ou non. Il nous faut également noter que ce 'trait' regroupe différents types d'expressivité. La valence émotionnelle, par exemple, peut être positive (p. ex. heureux) ou négative (p. ex. triste). Ceci pourrait mener à des effets moyennés dans notre analyse et masquer en partie le rôle joué par les différents aspects de l'expressivité.

Un résumé des différents sous-corpus se trouve en Tableau 1. Les

'traits' situationnels étant des continuums, certains corpus n'ont pas été classés (NC) si leur nature était ambiguë pour un 'trait' particulier. Le continuum entre parole lue et spontanée, par exemple, passe par la parole dite 'préparée' qui pourrait convenir aux conférences. Un autre exemple concerne les corpus de parole utilisés en synthèse vocale qui ne sont pas directement diffusés en tant que tels mais pourraient être utilisés pour des annonces publiques. Ils n'ont donc pas été classés pour le 'trait' média. Pour les interviews, seules les parties de l'interviewé ont été conservées. Le nombre de locuteurs par 'trait' est relativement équilibré, allant de 13 à 17 locuteurs pour une durée moyenne totale d'environ 2 heures.

Situation de communication (SC)	Lu	Média	Expressif
Commentaires sportifs	-	+	+
Conférence	NC	NC	-
Discours politique	+	+	NC
Interview	-	+	+
Explication d'itinéraire	-	-	-
Journal télévisé	+	+	NC
Parole expressive SV	+	NC	+
Parole neutre SV	+	NC	-
Lecture neutre	+	-	-
Narration	-	-	+

Tableau 1 – *Distribution des sous-corpus selon les trois 'traits' situationnels (les corpus SV ont été enregistrés afin d'entraîner un synthétiseur vocal).*

3. Méthodologie

Afin d'effectuer l'analyse phonétique du corpus, nous avons développé une méthodologie spécifique qui intègre des techniques de TAL. Pour chaque sous-corpus, la transcription orthographique est exploitée afin de produire sa phonétisation automatique à l'aide de l'outil TAL eLite (Colotte & Beaufort, 2005 ; Beaufort & Ruelle, 2006) développé pour la synthèse vocale. Cela permet de produire une phonétisation 'standard' du texte, correspondant à une parole neutre lue. Cette transcription est ensuite automatiquement alignée avec la transcription phonétique réellement prononcée par le locuteur (et qui a été manuellement vérifiée).

Cet alignement repose sur une version légèrement modifiée de la distance d'édition de Levenshtein (1966). Plusieurs adaptations ont été effectuées, notamment afin d'éviter les erreurs d'alignement lorsque deux modifications phonétiques surviennent dans un contexte phoné-

tique réduit :

1. Certaines substitutions de phonèmes ne sont pas pénalisées (e → ɛ, ø→œ, etc.) car elles pourraient résulter d'un choix subjectif de l'annotateur,

2. Les insertions et suppressions de silences ne sont pas pénalisées,

3. Un cout réduit a été assigné à la substitution de [i] en [j], qui est assez fréquente dans le corpus,

4. Une réduction du cout a également été effectuée pour les suppressions et insertions de schwas afin de favoriser cette modification.

Cette dernière modification a permis d'éviter des erreurs d'alignement telles que

b	E	l	Z	_	s	/	E	t	y	n
b	E	l	Z	/	@	_	s	t	y	n

au lieu de

b	E	l	Z	/	_	s	E	t	y	n
b	E	l	Z	@	_	s	/	t	y	n

Afin de récupérer l'alignement, la matrice produite par l'algorithme est parcourue en sens inverse. Toutes les modifications sont ensuite stockées selon leur type (insertion, suppression ou substitution).

Pour éviter les potentielles erreurs de phonétisation générées par eLite, les fichiers sons contenant des nombres (écrits en chiffres) ont été supprimés. Les noms propres étant très fréquents dans les commentaires sportifs, la suppression des fichiers correspondants aurait significativement réduit la taille de ce sous-corpus. Nous avons dès lors décidé de considérer uniquement les modifications de phonèmes ne survenant ni sur les noms propres, ni sur une syllabe précédant ou suivant un nom propre.

L'utilisation d'une phonétique automatiquement produite par un module TAL a pour avantage de permettre une comparaison facile de la prononciation du corpus avec une prononciation dite 'standard'. Cette dernière prend en compte la majorité des variations phonétiques obligatoires telles que les élisions ou liaisons dues au contexte linguistique. Ceci nous permet de fournir une analyse plus précise que celle résultant d'une simple comparaison avec un dictionnaire phonétisé, tout en étant complètement automatique.

4. Analyse phonétique et prosodique

4.1. Analyse des variations phonétiques

Dans cette section, nous considérons tout d'abord la proportion globale de variations phonétiques pour chaque 'trait' situationnel (4.1.1.).

Nous nous concentrons ensuite sur l'analyse de quatres variations phonétiques qui ont été déterminées qualitativement comme étant les plus fréquentes dans notre corpus : l'élision du schwa (4.1.2.), l'insertion de schwas finaux (4.1.3.), les élisions de liquides finales (4.1.4.) et les liaisons (4.1.5.).

4.1.1. Proportion globale de changements phonétiques

Les variations phonétiques sont ici analysées en comparant la phonétisation standard produite par notre module TAL et la véritable prononciation réalisée par le locuteur. La proportion globale de changements phonétiques est calculée comme étant le nombre total de modifications (suppressions, insertions ou substitutions), divisé par le nombre maximal de caractères, c'est-à-dire le nombre de caractères de la plus longue des deux transcriptions. La significativité statistique des résultats est calculée à l'aide de t-tests unilatéraux ou de tests de Wilcoxon en fonction de la normalité de la variable.

Le Tableau 2 montre des différences significatives concernant le nombre total de modifications phonétiques pour les dimensions média et lu (avec respectivement p=0.043 and p=1.2e-05). Ceci indique que la parole spontanée et non-média diffère fortement de ce que produit un module TAL générique. Au contraire, les corpus de parole lue exploités en synthèse de la parole contiennent, en moyenne, seulement 1.31% de modifications phonétiques. Cette constatation suggère le fait que, bien qu'un module TAL produise une phonétisation correcte pour de la parole neutre lue, il nécessite des modifications non-négligeables afin de produire de la parole spontanée ou non-média. Enfin, il nous faut noter que le nombre de modifications phonétiques ne semble pas être influencé par le 'trait' expressif / non expressif. Ce constat pourrait cependant s'expliquer par l'hétérogénéité de ce 'trait' qui regroupe notamment des émotions de valence opposée.

Les sections suivantes se concentrent sur des phénomènes phonétiques typiques aux différents 'traits'.

4.1.2. Elision du schwa

L'élision du schwa est l'une des variations phonétiques les plus complexes en français. Certains schwas peuvent ainsi être prononcés ou non au milieu ou à la fin du mot. Notre analyse exclut ici les schwas finaux qui peuvent davantage être considérés comme des phénomènes de liaison. Le pourcentage de schwas élidés est ici calculé comme le nombre de suppressions de schwas, en milieu de mot, divisé par le nombre to-

	Tout changement		Elision de [l] dans 'il'		Elision dans obstruante+ liquide		Liaison	
'trait' situationnel	+	-	+	-	+	-	+	-
Lu	1.78%	4.25%	8.33%	76.56%	3.65%	50.34%	59.33%	42.15%
Média	2.93%	4.11%	51.87%	96.67%	18.53%	49.99%	54.52%	44.77%
Expressif	3.57%	2.92%	52.53%	34.72%	33.54%	21.74%	44.53%	50.27%

Tableau 2 – *Résumé des changements phonétiques pour les trois 'traits' situationnels.*

tal de schwas en milieu de mot. La Figure 1 montre le rôle significatif joué par les distinctions entre parole lue et spontanée ou média et non média (avec respectivement p=0.005 and p=1.7e-04). Ceci indique que plus de schwas sont élidés en parole spontanée, ce qui confirme les résultats d'études antérieures (Fougeron, Goldman, & Frauenfelder, 2001 ; Hambye, 2005 ; Hansen, 1991). Il est intéressant de souligner que plus de schwas sont également élidés en parole non-média, ce qui s'explique notamment par le fait que la parole média est généralement associée à un niveau de langue plus élevé qui est corrélé avec des taux d'élisions plus faibles (Warnant, 1996). Nous observons cependant un taux important de variation inter-locuteurs, similairement à Burki et al. (2011).

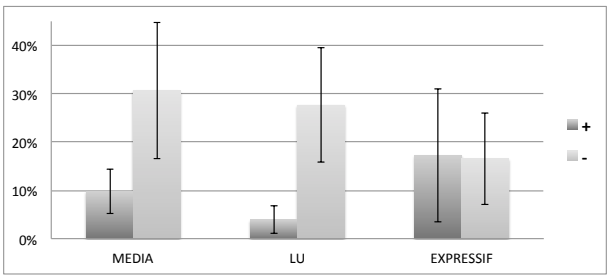

Figure 1 – *Pourcentage de schwas élidés en milieu de mot avec leur intervalle de confiance à 95%.*

4.1.3. Schwas finaux

Nous analysons ici l'insertion d'un schwa final en fin de mot (p. ex. *match* prononcé [matʃə̲]) (Hansen, 1991, 1994 ; Hansen & Hansen, 2003). Notre analyse se concentre ici sur les schwas survenant en fin de mot sans -e final que nous appelleront 'schwas épenthétiques'. Candea (2002) montre que leur fréquence a augmenté durant les dernières décennies et que leur apparition n'est plus dictée par le contexte rythmique ou pho-

nétique. Bien qu'ils aient longtemps été considérés comme une marque de discours informel, cet aspect sociolinguistique disparait peu à peu.

Il est intéressant de remarquer que l'apparition du schwa épenthétique est significativement plus fréquente dans la parole média comparé au non média, comme le montre la Figure 2 (p=0.013). Ceci corrobore les résultats de Roekhaut et al. (2010) qui indiquent un taux supérieur de prononciation de schwa finaux (tous mots confondus) dans les journaux radiophoniques et les discours politiques en comparaison à la parole conversationnelle. Ce taux est également significativement plus élevé dans la parole spontanée et expressive (respectivement p=0.019 et p=0.008), ce qui est très probablement du à leur haute fréquence d'apparition dans les commentaires sportifs. Une grande variabilité interlocuteurs est cependant observée.

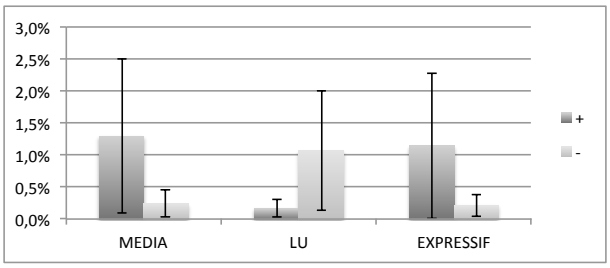

Figure 2 – *Pourcentage de mots, sans -e final, prononcés avec un schwa final et leur intervalle de confiance à 95%.*

4.1.4. Elisions de liquides finales

L'analyse de l'alignement entre prononciation réelle et prédite par eLite a souligné le fait que le pronom 'il' est souvent prononcé [i], avec élision du 'l' final. Ce phénomène survenant quasiment exclusivement avant un phonème consonantique, nous avons donc étudié son apparition dans ce contexte spécifique (voir Tableau 2). Seuls les sous-corpus contenant au moins 3 occurrences du pronom dans ce contexte ont été conservés. Les 'traits' média et lu sont significatifs, avec des taux d'élision plus élevés en parole spontanée et non-média (respectivement p= 6.1529e-05 and p=0.002) ce qui corrobore les résultats obtenus pour l'élision du schwa en Section 4.1.2..

Un second type d'élision de phonème consonantique concerne l'élision de la liquide finale lorsqu'elle est précédée d'une obstruante (p. ex. "peut-être" prononcé [pøtɛt]) (Reuse, 1987). Une première analyse qualitative du phénomène indique que cette variation dépend forte-

ment du contexte phonétique. La liquide est presque toujours prononcée si elle est suivie par une voyelle. Au contraire, elle peut être omise si elle est suivie par une consonne. Notre analyse s'est concentrée sur ce contexte et dans des sous-corpus contenant plus de 3 occurrences d'un tel contexte phonétique. Le Tableau 2 montre, ici aussi, que les corpus spontanés et non-média témoignent d'un pourcentage plus élevé d'élisions de la liquide comparé à la parole lue et média (respectivement p=1.1e-05 et p=0.04).

4.1.5. Liaisons

Nous avons défini les contextes potentiels de liaison comme étant les mots se terminant par une consonne de liaison en français / t, n, z, R, p / et suivis par une voyelle, similairement à d'autres études (Fougeron, Goldman, & Frauenfelder, 2001 ; Fougeron, Goldman, Dart, et al., 2001 ; Mareuil, Adda-Decker, & Gendner, 2003). Il est important de souligner ici que ces liaisons potentielles ne correspondent pas à ce qu'on appelle les 'liaisons optionnelles'. Toutes les liaisons sont donc considérées, qu'elles soient obligatoires, facultatives ou interdites. Le Tableau 2 montre que la parole lue contient un pourcentage significativement plus élevé de liaisons (p=4.6396e-05) ce qui confirme les résultats d'études existantes (Fougeron, Goldman, & Frauenfelder, 2001 ; Fougeron, Goldman, Dart, et al., 2001 ; Lucci, 1983). Il est intéressant de noter que la parole média témoigne également de plus d'élisions, même si la différence n'est pas significative. Ceci s'explique peut-être par le fait que la parole lue et média est généralement plus formelle, la parole dite 'soutenue' étant caractérisée par des taux de liaison plus élevés (Argod-Dutard, 1996 ; Warnant, 1996).

4.2. Prosodie : Corrélations entre caractéristiques rythmiques et phonétiques ?

Il a été montré que le rythme, et le débit de parole en particulier, ont tendance à être corrélés avec l'élision du schwa, la parole rapide contenant davantage de schwas élidés (Lacheret-Dujour, 1991). Cette section analyse différentes caractéristiques rythmiques afin de décrire leur lien avec les 'traits' situationnels et d'évaluer leur corrélation avec les différentes variations phonétiques précédemment décrites. Notre analyse se concentre sur trois mesures rythmiques : le débit d'articulation, la durée moyenne des unités interpausales (UIP) et la proportion de syllabes proéminentes. La significativité statistique des résultats est calculée à l'aide de t-tests unilatéraux ou de tests de Wilcoxon en fonction de la

normalité de la variable. Les corrélations sont évaluées à l'aide du coefficient de Spearman.

L'analyse du débit d'articulation[1] montre des différences significatives entre la parole lue et la parole spontanée, avec un débit d'articulation plus bas en parole spontanée (p=0.019). Ceci peut s'expliquer par la présence de syllabes allongées dues aux hésitations. Tout comme (Simon et al., 2009), nous observons également un pourcentage d'articulation significativement plus faible pour la parole spontanée (p=0.049). Cependant, contrairement aux études existantes (p. ex. (Lacheret-Dujour, 1991)), nous n'observons pas de corrélation entre le débit de parole et l'élision du schwa ($|Rho| < 0.09$), ou quelconque autre variation phonétique. Nous avons montré, au contraire, que les élisions sont plus fréquentes en parole spontanée, la parole spontanée étant caractérisée par un débit d'articulation plus faible. Cette différence peut s'expliquer par le fait que la majorité des études existantes se concentrent sur une tâche spécifique (p. ex. la lecture d'un texte) dans laquelle seul le débit d'articulation est modifié afin d'observer la fréquence des élisions. Dans notre corpus, de nombreux facteurs influencent ce débit de parole (p. ex. les hésitations, la situation de communication, les particularités idiosyncratiques, etc.).

La durée moyenne des UIP s'avère être significativement plus longue en parole spontanée comparée à la parole lue (p=0.04). Ceci semble indiquer que la parole spontanée contient moins de silences mais que ceux-ci sont plus longs. Une explication possible est que les pauses courtes sont rarement silencieuses en parole spontanée et sont souvent remplies par des marqueurs d'hésitation.

Enfin, le pourcentage de syllabes proéminentes est ici calculé à l'aide de Prosoprom (Goldman, Avanzi, Lacheret-Dujour, Simon, & Auchlin, 2007), un algorithme automatique qui permet de détecter les syllabes proéminentes sur base acoustique. Seule la dimension média semble jouer un rôle dans cette mesure, la parole média contenant plus de syllabes proéminentes (voir Figure 3). Cette distinction est d'ailleurs encore plus marquée lorsque l'on considère uniquement les accents initiaux, la parole média contenant 19.1 % de syllabes initiales proéminentes contre 16.4 % en parole non média. Ceci confirme les résultats de Goldman et al. (2009). Cependant, la variabilité inter-locuteurs est

[1] Nous nous concentrons ici sur le débit d'articulation, c'est-à-dire le débit de parole (nombre de syllabes par seconde) en excluant les silences, les différents sous-corpus témoignant de différentes densités de silence.

assez élevée et rend cette différence non significative (p=0.07).

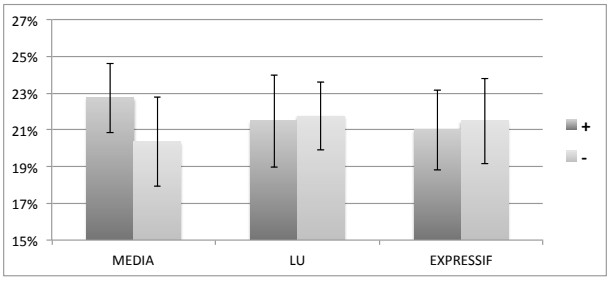

Figure 3 – *Pourcentage de syllabes proéminentes et leur intervalle de confiance à 95%.*

La durée des UIP est modérément corrélée aux élisions de schwas et du 'l' dans le pronom 'il' (avec respectivement $|Rho| = 0.59$ et $|Rho| = 0.43$). Ceci indique que plus d'élisions sont réalisées lorsque de plus grands segments de parole sont prononcés sans pause. Cette constatation n'est cependant pas valable pour l'élision de la liquide qui suit une obstruante. Enfin, il est intéressant de noter qu'aucune corrélation n'est attestée entre le pourcentage de proéminences et les différentes variations phonétique. Globalement, les corrélations entre variations phonétiques et rythmiques sont donc assez faibles. Ceci tend à indiquer que les variations phonétiques dépendent plutôt directement du 'trait' situationnel plutôt que des caractéristiques rythmiques de ce 'trait'. Notons également que, tout comme pour les variations phonétiques, le 'trait' expressif n'est pas caractérisé par des paramètres prosodiques particuliers.

5. Conclusion et perspectives

Alors que la synthèse vocale de différents styles de parole et émotions fait à présent l'objet de nombreuses recherches, peu d'études prennent en compte les variations phonétiques potentielles en fonction de la situation de communication visée (expressif, spontané, etc.). Cet article a proposé une analyse d'un large corpus en français afin d'évaluer l'influence exercée par trois 'traits' situationnels : lu / spontané, média/non-média et expressif / non-expressif. Nous avons tout d'abord montré que la parole spontanée et non média témoigne d'un pourcentage élevé de variations phonétiques en comparaison à une parole lue standard telle que prononcée par un synthétiseur vocal. En ce qui concerne les différents phénomènes, nous avons montré que la parole spontanée est caractérisée par un taux plus élevé d'élisions et moins de liaisons, ce

qui confirme les résultats d'études existantes. La parole média suit généralement les même tendances phonétiques que la parole lue, ce qui peut s'expliquer par le niveau de langue plus élevé de ce type de parole, contrairement à la parole spontanée ou non média. Cependant, elle contient beaucoup plus de schwas épenthétiques, phénomène qui semble être tout particulièrement caractéristique des commentaires sportifs. Enfin, il est intéressant de noter que le 'trait' expressif n'est pas associé à des caractéristiques phonétiques particulières. La diversité des corpus dans ce 'trait' (p. ex. émotions avec différentes valences) devrait être analysée plus en détail afin d'évaluer le rôle joué par les différents aspects de l'expressivité.

Dans un second temps, notre analyse s'est tournée vers les caractéristiques rythmiques des différents 'traits'. Cette analyse a mis en exergue un débit d'articulation plus élevé ainsi que des unités inter-pausales plus courtes pour la parole lue. De plus grandes proportions de proéminences ont également été observées pour la parole média. Une fois de plus, le 'trait' expressif n'a pu être associé à aucune caractéristique rythmique particulière. Enfin, des niveaux de corrélation bas ont été observés entre les paramètres rythmiques et les variations phonétiques, à l'exception d'une corrélation modérée entre la durée des unités inter-pausales et certains types d'élisions. Ceci indique que les variations phonétiques dépendent essentiellement du 'trait' situationnel (lu / spontané et média / non média) et non des caractéristiques rythmiques de ce 'trait'.

Notre recherche se tourne à présent sur l'analyse perceptive de ces changements phonétiques afin d'évaluer s'il s'agit simplement de variantes possibles ou si leur génération permettrait d'améliorer la crédibilité du message. Les variations phonétiques seront ensuite intégrées dans la synthèse vocale par modèles de Markov cachés (HMM) en fonction de la situation de communication visée.

6. Remerciements

Les auteurs sont soutenus par le FNRS. Ce projet est partiellement financé par le projet Wist 3 SPORTIC de la région wallonne. Les auteurs remercient également J.-P. Goldman pour ses conseils avisés.

Bibliographie

Argod-Dutard, F. (1996). *Eléments de phonétique appliquée.* Paris : Armand Colin.

Avanzi, M., Simon, A. C., Goldman, J. P., & Auchlin, A. (2010). C-PROM. An annotated corpus for French prominence studies. In *Actes de Proso-*

dic Prominence : Perceptual and Automatic Identification, Speech Prosody 2010 Workshop. Chicago (USA). Disponible sur http://speechprosody2010.illinois.edu/papers/102005.pdf

Beaufort, R., & Ruelle, A. (2006). eLite : système de synthèse de la parole à orientation linguistiques. In *Journées d'étude sur la parole (JEP).*

Brognaux, S., Picart, B., & Drugman, T. (2013). A new prosody annotation protocol for live sports commentaries. In F. Bimbot et al. (Éds), *Actes de Interspeech 2013* (p. 1554-1558). Lyon (France). Disponible sur http://tcts.fpms.ac.be/publications/papers/2013/interspeech2013_proso_sbbptd.pdf

Brognaux, S., Roekhaut, S., Drugman, T., & Beaufort, R. (2012). Train&Align : A new online tool for automatic phonetic alignments. In *Actes du IEEE Workshop on Spoken Language Technologies (SLT).* Miami (USA). Disponible sur http://cental.fltr.ucl.ac.be/train_and_align/

Burki, A., Ernestus, M., Gendrot, C., Fougeron, C., & Frauenfelder, U. H. (2011). What affects the presence versus absence of schwa and its duration : A corpus analysis of French connected speech. *The journal of the Acoustical Society of America*, *130 (6)*, 3980-3991.

Candea, M. (2002). Le e d'appui parisien : statut actuel et progression. In *Actes des XXIVe Journées d'Etudes sur la Parole (JEP).* Nancy (France). Disponible sur http://www.ilpga.univ-paris3.fr/pages-personnelles/maria_candea/candea-jep2002.PDF

Colotte, V., & Beaufort, R. (2005). Linguistic features weighting for a text-to-speech system without prosody model. In *Actes de Interspeech 2005* (p. 2549-2552). Lisbonne (Portugal). Disponible sur http://www.multitel.be/uploaded/publications/pub6_p1689_colotte-beaufort.pdf

Fougeron, C., Goldman, J.-P., Dart, A., Gulat, L., & Jeager, C. (2001). Influence de facteurs stylistiques, syntaxiques et lexicaux sur la réalisation de la liaison en français. In *Actes de la 8ème conférence Traitement Automatique des Langues Naturelles (TALN)* (p. 173-182). Tours (France). Disponible sur http://sites.univ-provence.fr/veronis/Atala/TALN/pdf/art15_p173_182.pdf

Fougeron, C., Goldman, J.-P., & Frauenfelder, U. H. (2001). Liaison and schwa deletion in French : an effect of lexical frequency and competition ? In P. Dasgaard, B. Lindberg, H. Benner, & Z. Tan (Éds), *Actes de Eurospeech 2001* (p. 639-642). Aalborg (Danemark). Disponible sur http://perso.telecom-paristech.fr/~chollet/Biblio/Congres/Audio/Eurospeech01/CDROM/papers/page639.pdf

Goldman, J.-P. (2011). EasyAlign : An automatic phonetic alignment tool under Praat. In *Actes de Interspeech 2011* (p. 3233-3236). Florence (Italie). Disponible sur http://www.isca-speech.org/archive/interspeech_2011/i11_3233.html

Goldman, J. P., Auchlin, A., & Simon, A. C. (2009). Discrimination de styles de parole par analyse prosodique semi-automatique. In *Actes de Interfaces Discours Prosodie 2009 (IDP)* (p. 207-221). Paris (France). Disponible sur http://makino.linguist.jussieu.fr/idp09/docs/IDP_actes/Articles/Goldman.pdf

Goldman, J.-P., Avanzi, M., Lacheret-Dujour, A., Simon, A. C., & Auchlin, A. (2007). A methodology for the automatic detection of perceived prominent syllables in spoken French. In *Actes de Interspeech 2007* (p. 98-101). Anvers (Belgique). Disponible sur http://www.isca-speech.org/archive/interspeech_2007/i07_0098.html

Hambye, F. (2005). *La prononciation du français contemporain en Belgique. variations, normes et identités.* Thèse de doctorat non publiée, Université catholique de Louvain, Belgique. Disponible sur http://refef.crifpe.ca/document/these/HAMBYE.pdf

Hansen, A. (1991). The covariation of [schwa] with style in Parisian French : an empirical study of 'E caduc' and pre-pausal [schwa]. In *Actes du ESCA Workshop on Phonetics and Phonology of Speaking Styles* (p. 30 (1-7)). Barcelone (Espagne). Disponible sur http://20.210-193-52.unknown.qala.com.sg/archive_open/archive_papers/pposPst/pp91_030.pdf

Hansen, A. (1994). Etude du e caduc - stabilisation en cours et variations lexicales. *Journal of French Language Studies*, *4*, 25-54.

Hansen, A., & Hansen, M.-B. (2003). Le schwa prépausal et l'interaction. *Etudes Romanes*, *54*, 89-109. Disponible sur http://www.academia.edu/download/30994784/Schwa.pdf

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words : Evidence from reduction in lexical production. *Typological studies in language*, *45*, 229-254. Disponible sur http://www.stanford.edu/~jurafsky/cmu.pdf

Lacheret-Dujour, A. (1991). Phonological variations in read speech, reduction phenomena and speaker classes : do allophonic choices represent speaking style ? In *Actes du ESCA Workshop on Phonetics and Phonology of Speaking Styles* (p. 38 (1-10)). Barcelone (Espagne). Disponible sur http://www.isca-speech.org/archive_open/pposPst/pp91_038.html

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10 (8)*, 707-710. Disponible sur http://profs.sci.univr.it/~liptak/ALBioinfo/files/levenshtein66.pdf

Lucci, V. (1983). *Etude phonétique du français contemporain à travers la variation situationnelle (débit, rythme, accent, intonation, e muet, liaisons, phonèmes)*. Publications de l'Université des Langues et Lettres de Grenoble.

Mareuil, P. Boula de, Adda-Decker, M., & Gendner, V. (2003). Liaisons

in French : a corpus-based study using morpho-syntactic information. In *Actes du 15th International Congress of Phonetic Sciences* (p. 1329-1332). Barcelone (Espagne). Disponible sur ftp://tlp.limsi.fr/public/ICPhSliaison.pdf

Martinet, A. (1971). *La prononciation du français contemporain*. Librairie Droz.

Picart, B., Brognaux, S., & Drugman, T. (2013). HMM-based speech synthesis of live sports commentaries : Integration of a two-layer prosody annotation. In *Actes du 8th ISCA Speech Synthesis Workshop (SSW8)* (p. 19-24). Barcelone (Espagne). Disponible sur http://ssw8.talp.cat/papers/ssw8_OS1-4_Picart.pdf

Picart, B., Drugman, T., & Dutoit, T. (2010). Analysis and synthesis of hypo and hyperarticulated speech. In *Actes du 7th Speech Synthesis Workshop (SSW7)* (p. 270-275). Kyoto (Japon). Disponible sur http://20.210-193-52.unknown.qala.com.sg/archive/ssw7/papers/ssw7_270.pdf

Pršir, T., Goldman, J. P., & Auchlin, A. (2013). Variation prosodique situationnelle : étude sur corpus de huit phonogenres en français. In P. Mertens & A. C. Simon (Éds), *Actes de Interfaces Discours Prosodie 2013 (IDP)* (p. 107-111). Leuvent (Belgique). Disponible sur http://wwwling.arts.kuleuven.be/franitalco/idp2013/Proceedings.html

Reuse, J. W. de. (1987). La phonologie du français de la région de Charleroi (Belgique) et ses rapports avec le wallon. *La linguistique*, *23*, 99-115. Disponible sur http://www.jstor.org/stable/30248975

Roekhaut, S., Goldman, J.-P., & Simon, A. C. (2010). A model for varying speaking style in TTS systems. In *Actes de Speech Prosody 2010* (p. 1-4). Chicago (USA). Disponible sur http://speechprosody2010.illinois.edu/papers/100096.pdf

Simon, A. C., Auchlin, A., Avanzi, M., & Goldman, J. P. (2009). Les phonostyles : une description prosodique des styles de parole en français. In P. Lang (Éd.), *Les voix des Français. en parlant, en écrivant vol. 2* (p. 71-88). Berne : Abecassis, M. & G. Ledegen.

Tsuzuki, R., Zen, H., Tokuda, K., Kitamura, T., Bulut, M., & Narayanan, S. (2004). Constructing emotional speech synthesizers with limited speech database. In *Actes de Interspeech 2004* (p. 1185-1188). Jeju Island (Korea). Disponible sur http://www.isca-speech.org/archive/interspeech_2004/i04_1185.html

Warnant, L. (1996). *Orthographe et prononciation en français. les 12000 mots qui ne se prononcent pas comme ils s'écrivent.* De Boeck-Duculot.

Yamagishi, J., Onishi, K., Masuko, T., & Kobayashi, T. (2005). Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, *88*(3), 502-509.